



3D Object Modeling and Recognition Using Affine-Invariant Patches and Multi-View Spatial Constraints

Fredrick Rothganger, Svetlana Lazebnik, Cordelia Schmid, Jean Ponce

► To cite this version:

Fredrick Rothganger, Svetlana Lazebnik, Cordelia Schmid, Jean Ponce. 3D Object Modeling and Recognition Using Affine-Invariant Patches and Multi-View Spatial Constraints. International Conference on Computer Vision & Pattern Recognition (CVPR '03), Jun 2003, Madison, United States. pp.272–277, 10.1109/CVPR.2003.1211480 . inria-00548224

HAL Id: inria-00548224

<https://inria.hal.science/inria-00548224>

Submitted on 20 Dec 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

3D Object Modeling and Recognition Using Affine-Invariant Patches and Multi-View Spatial Constraints

Fredrick Rothganger¹, Svetlana Lazebnik¹, Cordelia Schmid², and Jean Ponce¹

¹ Department of Computer Science and Beckman Institute

University of Illinois at Urbana-Champaign; Urbana, IL 61801, USA

² INRIA Rhône-Alpes; 665, Avenue de l'Europe; 38330 Montbonnot, France

Abstract: *This paper presents a novel representation for three-dimensional objects in terms of affine-invariant image patches and their spatial relationships. Multi-view constraints associated with groups of patches are combined with a normalized representation of their appearance to guide matching and reconstruction, allowing the acquisition of true three-dimensional affine and Euclidean models from multiple images and their recognition in a single photograph taken from an arbitrary viewpoint. The proposed approach does not require a separate segmentation stage and is applicable to cluttered scenes. Preliminary modeling and recognition results are presented.*

1 Introduction

This paper addresses the problem of recognizing three-dimensional (3D) objects in photographs. Traditional feature-based geometric approaches to this problem, for example alignment and interpretation trees [6, 8], enumerate all triples of image features before pose consistency constraints can be used to confirm or discard competing match hypotheses. Appearance-based techniques, on the other hand, use rich local descriptions of the image brightness pattern to select a relatively small set of promising potential matches before using geometric consistency constraints to retain the correct ones. However, they normally either require storing a large number of views for each object (e.g., [15, 17, 20]), or limiting the range of admissible viewpoints (e.g., [21, 25]).

Viewpoint invariants (or *invariants* for short) provide a natural indexing mechanism for object recognition tasks. Unfortunately, although planar objects and certain simple shapes (e.g., bilaterally symmetric ones) admit invariants, general 3D shapes do not [2], which is the main reason why invariants have fallen out of favor after an intense flurry of activity in the early 1990s [13, 14]. We propose in this paper to revisit invariants as a *local* object description: Indeed, although smooth surfaces are almost never planar in the large, they are *always* planar in the small—that is, sufficiently small surface patches can always be thought of as being comprised of coplanar points. The surface of a solid

can thus be represented by a collection of small patches, their invariants, and a description of their 3D spatial relationship.

We propose such a surface representation in terms of the affine-invariant patches introduced by Mikolajczyk and Schmid [12] and geometric consistency constraints related to the multi-view geometry studied in the structure-from-motion literature [3, 5, 23]. The detection and representation of affine-invariant patches is discussed in Section 2. We show in Section 3 that it is possible to exploit the multi-view geometry of affine projection to impose effective pose-consistency constraints on matching patches (see [10, 19, 22, 24] for related work). Thus, different views of the same scene can be matched by checking whether groups of potential correspondences found by correlation are geometrically consistent. This matching strategy is used in modeling tasks, where matches found in pairs of successive (unregistered) images of the same object are used to create a global 3D (affine or Euclidean) model of this object. We show in Section 4 that multi-view consistency constraints imposed on potential correspondences between these 3D models and image patches can also be used to identify 3D objects in photographs taken from arbitrary viewpoints. In practice, modeling is done in controlled situations with little or no clutter; the stronger consistency constraints associated with 3D models make up for the presence of significant clutter in object recognition tasks, avoiding the need for a separate segmentation stage. The proposed approach to 3D object modeling and recognition has been implemented, and preliminary experiments are presented in Section 5. We conclude in Section 6 with a brief discussion of its promise and limitations.

2 Affine-Invariant Patches

Operators capable of finding rotation- [20], scale- [9, 11] and affine-invariant [1, 12, 18, 24] image descriptors in the neighborhood of salient image features (“interest points” [4]) have recently been proposed in the context of wide-baseline stereo matching and image retrieval. In this paper, we use an implementation of the affine-invariant region detector developed by Mikolajczyk and Schmid [12] for low-level im-

age description. In this approach, the dependency of an image patch’s appearance on affine transformations of the image plane is eliminated by an iterative rectification process based on using (a) the second-moment matrix computed in the neighborhood of a point to normalize the shape of the corresponding image patch in an affine-invariant manner; (b) the local extrema of the normalized Laplacian over scale to determine the characteristic scale of the local brightness pattern; (c) an affine-adapted Harris detector to determine the patch location; and (d) image gradient information to eliminate any remaining rotational ambiguity.

The output of the interest point detection/rectification process is a set of image patches in the shape of parallelograms, together with the corresponding affine *rectifying transformations*. The transformation \mathcal{R} associated with each patch maps the corresponding parallelogram onto a square with unit edge half-length centered at the origin (Figure 1). The rectified patch is a *normalized* representation of the local surface appearance that is invariant under planar affine transformations. We will assume from now on an affine—that is, orthographic, weak-perspective, or paraperspective—projection model (the full perspective case will be briefly discussed in Section 6). Under this model, our normalized appearance representation is invariant under arbitrary changes in viewpoint. For Lambertian patches and distant light sources, it can also be made invariant to changes in illumination (ignoring shadows) by subtracting the mean patch intensity from each pixel value and normalizing the sum of squared intensity values to one (or equivalently using *normalized* correlation to compare patches).

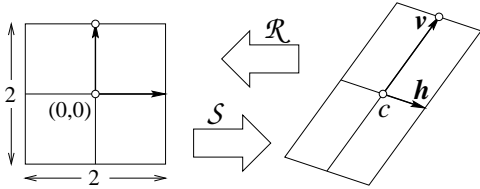


Figure 1: Geometric interpretation of the rectification matrix \mathcal{R} and its inverse \mathcal{S} .

The rectifying transformation associated with a planar patch and its inverse can be represented by two 2×3 matrices \mathcal{R} and \mathcal{S} that map homogeneous (affine) plane coordinates onto non-homogeneous ones. These transformations will play a fundamental role in the rest of this paper. For the time being, let us note that the columns vectors of the matrix \mathcal{S} admit a simple geometric interpretation: Since they are respectively the images of the vectors $(1, 0, 0)^T$, $(0, 1, 0)^T$, and $(0, 0, 1)^T$ under that mapping, the third column \mathbf{c} of \mathcal{S} is the (non-homogeneous) coordinate vector of the patch center c , and its first two columns \mathbf{h} and \mathbf{v} are respectively the (non-homogeneous) coordinate vectors of the “horizontal” and “vertical” vectors joining c to the sides of the patch (Figure 1). These two vectors can also be interpreted as the po-

sitions of the points, dubbed *normalized side points* in the sequel, where the “horizontal” and “vertical” axes of a copy of the image patch placed at the origin pierce its right and top side.

In particular (and not surprisingly), a match between $m \geq 2$ images of the same affine invariant patches contains *exactly* the same information as a match between m triples of points. It is thus clear that all the machinery of structure from motion [3, 5, 23] and pose estimation [6, 8] from point matches can be exploited in our modeling and object recognition tasks. Reasoning in terms of multi-view constraints associated with the matrix \mathcal{S} will provide in this paper a unified and convenient representation for all stages of both tasks, but one should always keep in mind the simple geometric interpretation of the matrix \mathcal{S} and the deeply rooted relationship between these constraints and those used in motion analysis and pose estimation. We will come back to this issue in Section 6.

3 3D Object Modeling

In this section, we exploit the multi-view geometry of affine projection to impose effective pose-consistency constraints on matching patches. This allows us to match different views of the same scene by checking whether groups of potential correspondences found by correlation are geometrically consistent. Matches found in pairs of successive (unregistered) images of the same object are finally stitched into a global 3D (affine or Euclidean) model of this object.

3.1 Matching Constraints

Let us assume for the time being that we are given n patches observed in m images, together with the corresponding 2×3 matrices \mathcal{R}_{ij} and \mathcal{S}_{ij} defined as in Section 2 for $i = 1, \dots, m$ and $j = 1, \dots, n$ (i and j serving respectively as image and patch indices). We use these transformations to derive in this section a set of geometric and algebraic constraints that must be satisfied by matching image patches.

A rectified patch can be thought of as a *fictitious* view of the original surface patch (Figure 2), and the inverse mapping \mathcal{S}_{ij} can thus be decomposed into an *inverse projection* \mathcal{N}_j [3] that maps the rectified patch onto the corresponding surface patch, followed by a projection \mathcal{M}_i that maps that patch onto its projection into image number i , i.e., $\mathcal{S}_{ij} = \mathcal{M}_i \mathcal{N}_j$ for $i = 1, \dots, m$ and $j = 1, \dots, n$. (This is an affine instance of the characterization of homographies induced by planes given in Faugeras, Luong and Papadopoulos [3, Prop. 5.1].)

In particular, we can write

$$\hat{\mathcal{S}} \stackrel{\text{def}}{=} \begin{bmatrix} \mathcal{S}_{11} & \dots & \mathcal{S}_{1n} \\ \dots & \dots & \dots \\ \mathcal{S}_{m1} & \dots & \mathcal{S}_{mn} \end{bmatrix} = \begin{bmatrix} \mathcal{M}_1 \\ \vdots \\ \mathcal{M}_m \end{bmatrix} [\mathcal{N}_1 \quad \dots \quad \mathcal{N}_n],$$

and it follows that the $2m \times 3n$ matrix $\hat{\mathcal{S}}$ has at most rank 4.

We have not taken into account (so far) the form of the inverse projection matrix. As shown in the appendix, it can

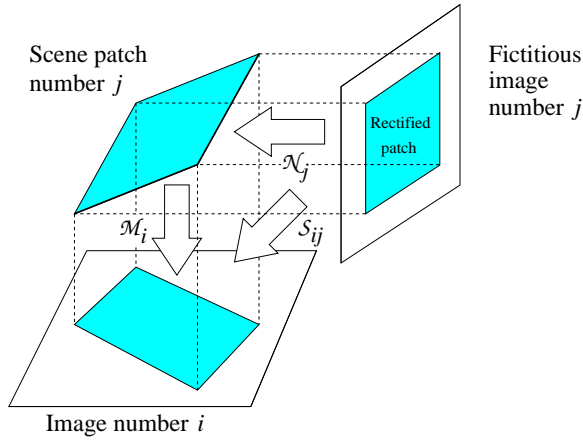


Figure 2: Geometric interpretation of the decomposition of the mapping \mathcal{S}_{ij} into the product of a projection matrix \mathcal{M}_i and an inverse projection matrix \mathcal{N}_j .

be written as

$$\mathcal{N}_j = \begin{bmatrix} \mathcal{B}_j \\ (0, 0, 1) \end{bmatrix},$$

where \mathcal{B}_j is a 3×3 matrix, and it satisfies the constraint $\mathcal{N}_j^T \mathbf{\Pi}_j = 0$, where $\mathbf{\Pi}_j$ is the coordinate vector of the plane Π_j that contains the patch. In addition, the columns of the matrix \mathcal{B}_j admit in our case a geometric interpretation related to that of the matrix \mathcal{S}_{ij} : Namely, the first two are the (non-homogeneous) coordinate vectors of the “horizontal” and “vertical” axes of the surface patch, and the third one is the (non-homogeneous) coordinate vector of its center \mathcal{C}_j (see appendix).

To account for the form of \mathcal{N}_j , we construct a reduced factorization of $\hat{\mathcal{S}}$ by picking, as in Tomasi and Kanade [23], the center of mass of the observed patches’ centers as the origin of the world coordinate system, and the center of mass of these points’ projections as the origin of every image coordinate system: In this case, the projection matrices reduce to $\mathcal{M}_i = [\mathcal{A}_i \quad \mathbf{0}]$, where \mathcal{A}_i is a 2×3 matrix, and $\mathcal{S}_{ij} = \mathcal{A}_i \mathcal{B}_j$. It follows that the reduced $2m \times 3n$ matrix

$$\hat{\mathcal{S}} = \hat{\mathcal{A}} \hat{\mathcal{B}}, \text{ where } \hat{\mathcal{A}} \stackrel{\text{def}}{=} \begin{bmatrix} \mathcal{A}_1 \\ \vdots \\ \mathcal{A}_m \end{bmatrix}, \quad \hat{\mathcal{B}} \stackrel{\text{def}}{=} [\mathcal{B}_1 \quad \dots \quad \mathcal{B}_n],$$

has at most rank 3, a fact that can be used as a matching constraint when at least two matches are visible in at least two views.

Alternatively, singular value decomposition can be used as in Tomasi and Kanade [23] to factorize $\hat{\mathcal{S}}$ and compute estimates of the matrices $\hat{\mathcal{A}}$ and $\hat{\mathcal{B}}$ that minimize the squared Frobenius norm of the matrix $\hat{\mathcal{S}} - \hat{\mathcal{A}} \hat{\mathcal{B}}$. The residual (normalized) Frobenius form $|\hat{\mathcal{S}} - \hat{\mathcal{A}} \hat{\mathcal{B}}| / \sqrt{3mn}$ of this matrix can be interpreted geometrically as the average distance (in pixels) between the center and normalized side points of the patches

observed in the image, and the center and normalized side points predicted from the recovered matrices $\hat{\mathcal{A}}$ and $\hat{\mathcal{B}}$.

3.2 Matching Strategy

Image matching requires two key ingredients: (a) a measure of appearance similarity between two images of the same patch, and (b) a measure of geometric consistency between n matches M_1, \dots, M_n established across m images (a match is an m -tuple of image patches). For the former we use normalized correlation between rectified patches. For the latter, we use the method described in the previous section to estimate (when $m, n \geq 2$) the matrices $\hat{\mathcal{A}}$ and $\hat{\mathcal{B}}$, and define $d(M_1, \dots, M_n) = |\hat{\mathcal{S}} - \hat{\mathcal{A}} \hat{\mathcal{B}}| / \sqrt{3mn}$ as a measure of inconsistency between the matches.

In our current implementation, we only match patches across pairs of images ($m = 2$), and follow a strategy similar to that used in the range data domain by Johnson and Hebert [7] with *spin images*. Given a patch in one image, we first select its N most promising matches in the second image based on normalized correlation of the rectified patches. We then find groups of consistent matches as follows: For each one of the p matches, we initialize the group G to that match M . We then find the match M' minimizing $d(G, M')$ (naturally defined as $d(M_1, \dots, M_k, M')$ when $G = (M_1, \dots, M_k)$). If $d(G, M') < T$, where T is a preset threshold, we add M' to G and continue. This results in the construction of p groups. Finally, we discard all groups smaller than some threshold Q . The remaining matches are judged to be correct. We then use estimated projection matrices to predict additional matches.

The implementation of this matching strategy is determined by the choice of the three thresholds N , T , and Q . In the experiments presented in Section 5 we have used $N = 5$, and only group matches with correlation above 0.9. We determine the other two thresholds from statistics on the data itself.

3.3 Constructing an Integrated Model

The matching strategy outlined in the previous section can be used in modeling tasks to match successive pairs of views of the same object. The matching process provides as a side benefit the affine structure of the scene: The planes Π_j are the zero eigenvectors of the matrices \mathcal{N}_j^T , and the points \mathcal{C}_j are the third columns of these matrices.

When some of the patches are only observed in some of the frames (the usual case), the data can be split into overlapping blocks of two or more frames, using all the patches visible in all images of the same block to run the factorization technique, then using the points common to overlapping blocks to register the successive reconstructions in a common frame. In principle, it is sufficient to have blocks that overlap by four points. Once all blocks are registered, the initial estimates of the variables \mathcal{M}_i and \mathcal{N}_j are refined by minimizing $\sum_{j=1}^n \sum_{i \in I_j} |\mathcal{S}_{ij} - \mathcal{M}_i \mathcal{N}_j|^2$, where I_j denotes the set of images where patch number j is visible. Given the reasonable guesses available from the initial registration, this

non-linear least-squares process only takes (in general) a few iterations to converge.

It is not possible to go from affine to Euclidean structure and motion from two views only. When three or more views are available, on the other hand, it is a simple matter to compute the corresponding Euclidean weak-perspective projection matrices (assuming the aspect-ratios are known) and recover the Euclidean structure [16, 23].

4 3D Object Recognition

We now assume that the technique described in Section 3 has been used to create a library of 3D object models, and address the problem of identifying instances of these models in a test image. As before, we start by deriving matching constraints between model and image patches before using these constraints to retain or discard groups of potential correspondences.

4.1 Matching Constraints

We assume in the rest of this section that $n \geq 2$ affine-invariant patches found in a test image have been putatively matched to n patches from a single object model, and derive consistency constraints that must be satisfied by these matches. Let us assume that we have the rectification matrices $\mathcal{S}_1, \dots, \mathcal{S}_n$ associated with the corresponding patches in the test image. As in Section 3, we can always pick the center of mass of the n patch centers in the test image as the origin of its coordinate system, and change the origin of the world coordinate system so that it coincides with the center of mass of their matches in the model.

With this convention, the projection matrix can be written as $\mathcal{M} = [\mathcal{A} \ 0]$ and we can write as before $\mathcal{S}_j = \mathcal{A}\mathcal{B}_j$. We have therefore $\mathcal{A} = \mathcal{S}_j\mathcal{B}_j^{-1}$ for $j = 1, \dots, n$. Note that the value of \mathcal{B}_j is available from the modeling stage *in the coordinate system attached to the model*. If \mathcal{C} denotes the (known) position of the center of mass of the patch centers in the original coordinate system, it is easy to see that the value of \mathcal{B}_j in the new coordinate frame is obtained by subtracting $\mathcal{C}(0, 0, 1)$ from its old value. We can now write

$$\tilde{\mathcal{B}}\mathcal{A}^T = \tilde{\mathcal{S}}, \text{ where } \tilde{\mathcal{B}} \stackrel{\text{def}}{=} \begin{bmatrix} \mathcal{B}_1^T \\ \dots \\ \mathcal{B}_n^T \end{bmatrix}, \tilde{\mathcal{S}} \stackrel{\text{def}}{=} \begin{bmatrix} \mathcal{S}_1^T \\ \dots \\ \mathcal{S}_n^T \end{bmatrix},$$

and the least-squares solution of this equation is $\mathcal{A}^T = \tilde{\mathcal{B}}^\dagger \tilde{\mathcal{S}}$, where $\tilde{\mathcal{B}}^\dagger$ denotes the pseudoinverse of $\tilde{\mathcal{B}}$. An appropriate measure of consistency in this case is (similar to the image matching case) $|\tilde{\mathcal{S}} - \tilde{\mathcal{B}}\mathcal{A}^T|/\sqrt{3n}$, that can once again be interpreted in terms of image distances measured in pixels.

4.2 Matching Strategy

Our current implementation uses the affine matching constraints derived in Section 4.1. Normalized correlation between rectified patches is once again used as a measure of appearance similarity between image patches. Note that a representative image patch must be chosen for each patch in

the model. We pick the patch with the largest characteristic scale. For geometric consistency, we use the method described in the previous section to estimate (when $n \geq 2$) the matrices \mathcal{A}^T and use $d'(M_1, \dots, M_n) \stackrel{\text{def}}{=} |\tilde{\mathcal{S}} - \tilde{\mathcal{B}}\mathcal{A}^T|/\sqrt{3n}$ as a measure of inconsistency between the matches. Except for these minor differences, the matching strategy remains the same as in the modeling case.

5 Implementation

We have implemented the approach proposed in Sections 3 and 4 and present preliminary modeling and recognition experiments. Since both the modeling and matching constraints are based on the inverse rectification matrices \mathcal{S}_{ij} , it is important to estimate these matrices as accurately as possible. Initial estimates provided by the affine-invariant region detector are improved as follows at every stage of the matching/reconstruction process. Assuming that patch number j is seen in m_j views, we can adjust the inverse rectification matrices \mathcal{S}_{ij} by maximizing the average normalized correlation between the $m_j(m_j - 1)/2$ pairs of (inversely rectified) images of this patch (see Figure 3).



Figure 3: Adjusting the rectifying transformations: Rectified patches associated with a match in three views before (top) and after applying the refinement process (bottom).

We have applied the modeling approach described in Section 3 to several different objects, four of which are shown in Figure 4. For each object, the figure shows one sample input picture from the set of input pictures. Each set contains an average of 16 input pictures. Below each input picture, the figure shows a rendering of the Euclidean model. The models are rather sparse, but one should keep in mind that they are intended for object recognition, not for image-based rendering applications.

Figure 5 shows the results of some recognition experiments. On top are the test pictures, with recognized patches marked. Some of the pictures contain instances of two different models. Below are the models rendered in their recovered poses. The average re-projection error is small in all cases, ranging from 2.8 pixels for the bear to 0.7 pixels for the salt, for 2200x1700 images.

6 Discussion

We have proposed in this paper to revisit invariants as a local object description that exploits the fact that smooth surfaces are always planar in the small. Combining this idea with the affine-invariant patches of Mikolajczyk and Schmid



Figure 5: Recognition experiments. (Top) Test pictures, with recognized patches marked. (Bottom) Models rendered in their recovered poses. (Left to right) Teddy bear, vase, salt and rubble together.



Figure 4: Object gallery. (Top) One of several input pictures for each object. (Bottom) Rendering of each model, not necessarily in same pose as input picture. (Left to right) Teddy bear, vase, plastic model of a pile of rubble, Morton Salt container.

[12] has allowed us to construct a normalized representation of local surface appearance that can be used to select promising matches in 3D object modeling and recognition tasks. We have used multi-view geometric constraints to represent the 3D surface structure in the large, retain groups of consistent matches, and reject incorrect ones. We believe that our preliminary experiments demonstrate the promise of the proposed approach.

The current implementation is limited to affine viewing conditions. As noted in Section 2, a match between $m \geq 2$ images of the same affine invariant patches is equivalent to a match between m triples of points, thus the machinery de-

veloped in the structure from motion [3, 5, 23] and pose estimation [6, 8] literature can in principle be used to extend our approach to the perspective case. Perhaps more interestingly, it might be possible to mix local affine constraints with global perspective/projective ones: Indeed, for patches whose relief is small compared to the distance separating them from the camera, the local projective distortions associated with the perspective projection process are normally negligible, and the rectifying transformations can thus be modeled as planar homographies that just happen to have an affine form. As shown in [3] for example, planar homographies can also be written as the composition of a (perspective) projection and an inverse projection, although this factorization is only defined up to an unknown scale factor, preventing the straightforward use of singular value decomposition techniques. Thus it would be interesting to combine affine and perspective/projective matching constraints in modeling and recognition tasks (as was done by Tuytelaars and Van Gool [24] in the image matching domain). Another problem that we feel has not been dealt with in a satisfactory manner is the combinatorics of the matching process: What is the best way (or at least a good way) to combine multiple views of multiple objects with some patches visible in some images but not in others? This remains an open question that we plan to address in the future.

Acknowledgments: This research was partially supported by the UIUC Campus Research Board, by the National Science Foundation under grant IRI-990709, by the CNRS-

Appendix: Inverse Projection Matrices

Consider a plane Π with coordinate vector Π in the world coordinate system. For any point in this plane we can write the affine projection in some image plane as $\mathbf{p} = \mathcal{M}\mathbf{P}$ and $\Pi \cdot \mathbf{P} = 0$. These two equations determine the homogeneous coordinate vector \mathbf{P} up to scale. To completely determine it, we can impose that its fourth coordinate be 1, and the corresponding equations become

$$\mathcal{M}_{\Pi}\mathbf{P} = \begin{bmatrix} \mathbf{p} \\ 0 \\ 1 \end{bmatrix}, \quad \text{where} \quad \mathcal{M}_{\Pi} = \begin{bmatrix} \mathcal{M} \\ \Pi^T \\ (0, 0, 0, 1) \end{bmatrix}.$$

Not surprisingly, \mathcal{M}_{Π} is an affine transformation matrix. So is its inverse, and if

$$\mathcal{M}_{\Pi}^{-1} = \begin{bmatrix} \mathbf{c}_1 & \mathbf{c}_2 & \mathbf{c}_3 & \mathbf{c}_4 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

we can write

$$\mathbf{P} = \mathcal{M}_{\Pi}^{\dagger} \begin{bmatrix} \mathbf{p} \\ 1 \end{bmatrix}, \quad \text{where} \quad \mathcal{M}_{\Pi}^{\dagger} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{c}_1 & \mathbf{c}_2 & \mathbf{c}_4 \\ 0 & 0 & 1 \end{bmatrix}.$$

The 4×3 matrix $\mathcal{M}_{\Pi}^{\dagger}$ is the *inverse projection matrix* [3] associated with the plane Π . Note that, for any point \mathbf{p} in the image plane, the point

$$\mathbf{P} = \mathcal{M}_{\Pi}^{\dagger} \begin{bmatrix} \mathbf{p} \\ 1 \end{bmatrix}$$

lies in the plane Π , thus $\Pi^T \mathbf{P} = 0$. Since this must be true for all points \mathbf{p} , we must have $\Pi^T \mathcal{M}_{\Pi}^{\dagger} = \mathbf{0}^T$.

The matrix \mathcal{N}_j used in this paper is simply $\mathcal{M}_{\Pi_j}^{(j)\dagger}$ where $\mathcal{M}^{(j)}$ is the matrix associated with the projection into the (fictitious) rectified image plane. Note that $\mathcal{M}^{(j)}$ maps the center C_j of patch number j onto the origin of the rectified image plane. It follows that the (non-homogeneous) coordinate vector of this point is

$$\mathbf{C}_j = \mathcal{N}_j \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

or, equivalently, that \mathbf{C}_j is the third column of the matrix \mathcal{N}_j . A similar reasoning shows that the “horizontal” and “vertical” axes of the patch are respectively the first and second columns of \mathcal{N}_j .

References

- [1] A. Baumberg. Reliable feature matching across widely separated views. In *Proc. CVPR*, pages 774–781, 2000.
- [2] J. B. Burns, R. S. Weiss, and E. M. Riseman. View variation of point-set and line-segment features. *IEEE Trans. PAMI*, 15(1):51–68, 1993.
- [3] O. Faugeras, Q.-T. Luong, and T. Papadopoulos. *The Geometry of Multiple Images*. MIT Press, 2001.
- [4] C. Harris and M. Stephens. A combined edge and corner detector. In *4th Alvey Vision Conference*, pages 189–192, 1988.
- [5] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2000.
- [6] D.P. Huttenlocher and S. Ullman. Object recognition using alignment. In *Proc. ICCV*, pages 102–111, 1987.
- [7] A.E. Johnson and M. Hebert. Surface matching for object recognition in complex three-dimensional scenes. *Image and Vision Computing*, 16:635–651, 1998.
- [8] D.G. Lowe. The viewpoint consistency constraint. *IJCV*, 1(1):57–72, 1987.
- [9] D.G. Lowe. Object recognition from local scale-invariant features. In *Proc. ICCV*, pages 1150–1157, 1999.
- [10] D.G. Lowe. Local Feature View Clustering for 3D Object Recognition In *Proc. CVPR*, I:682–688, 2001.
- [11] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Proc. ICCV*, pages 525–531, 2001.
- [12] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proc. ECCV*, vol. I, pages 128–142, 2002.
- [13] J.L. Mundy and A. Zisserman. *Geometric Invariance in Computer Vision*. MIT Press, 1992.
- [14] J.L. Mundy, A. Zisserman, and D. Forsyth. *Applications of Invariance in Computer Vision*, vol. 825, LNCS. Springer-Verlag, 1994.
- [15] H. Murase and S. Nayar. Visual learning and recognition of 3D objects from appearance. *IJCV*, 14(1):5–24, 1995.
- [16] J. Ponce. Metric upgrade of a projective reconstruction under the rectangular pixel assumption. In *Second SMILE Workshop*, pages 18–27, 2000.
- [17] A.R. Pope and D.G. Lowe. Probabilistic models of appearance for 3-D object recognition. *IJCV*, 40(2):149–167, 2000.
- [18] P. Pritchett and A. Zisserman. Wide baseline stereo matching. In *Proc. ICCV*, pages 754–760, 1998.
- [19] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or “How do I organize my holiday snaps?”. In *Proc. ECCV*, I:414–431, 2002.
- [20] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Trans. PAMI*, 19(5):530–535, 1997.
- [21] H. Schneiderman and T. Kanade. A statistical method for 3D object detection applied to faces and cars. In *Proc. CVPR*, 2000.
- [22] D. Tell and S. Carlsson. Wide baseline point matching using affine invariants computed from intensity profiles. In *Proc. ECCV*, pages 814–828, 2000.
- [23] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *IJCV*, 9(2):137–154, 1992.
- [24] T. Tuytelaars and L. Van Gool. Matching widely separated views based on affinely invariant neighborhoods. *IJCV*, 2002. Submitted.
- [25] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *Proc. ECCV*, 2000.